Compositional analysis of archaeological glasses

M.J. Baxter[1], C.C. Beardah[1], and I.C. Freestone[2]
[1]School of Biomedical and Natural Sciences, Nottingham Trent University, Clifton Campus, Nottingham NG11 8NS, UK
[2]Cardiff School of History and Archaeology, Cardiff University, Humanities Building, Colum Drive, Cardiff CF10 3EU, UK

Abstract

At CoDaWork'03 we presented work on the analysis of archaeological glass compositional data. Such data typically consist of geochemical compositions involving 10-12 variables and approximates completely compositional data if the main component, silica, is included. We suggested that what has been termed 'crude' principal component analysis (PCA) of standardized data often identified interpretable pattern in the data more readily than analyses based on log-ratio transformed data (LRA). The fundamental problem is that, in LRA, minor oxides with high relative variation, that may not be structure carrying, can dominate an analysis and obscure pattern associated with variables present at higher absolute levels. We investigate this further using subcompositional data relating to archaeological glasses found on Israeli sites. A simple model for glass-making is that it is based on a 'recipe' consisting of two 'ingredients', sand and a source of soda. Our analysis focuses on the sub-composition of components associated with the sand source. A 'crude' PCA of standardized data shows two clear compositional groups that can be interpreted in terms of different recipes being used at different periods, reflected in absolute differences in the composition. LRA analysis can be undertaken either by normalizing the data or defining a 'residual'. In either case, after some 'tuning', these groups are recovered. The results from the normalized LRA are differently interpreted as showing that the source of sand used to make the glass differed. These results are complementary. One relates to the recipe used. The other relates to the composition (and presumed sources) of one of the ingredients. It seems to be axiomatic in some expositions of LRA that statistical analysis of compositional data should focus on relative variation via the use of ratios. Our analysis suggests that absolute differences can also be informative.

Key words: Archaeometry, compositional data, glass, log-ratio analysis, principal components analysis.

1 Introduction

At CoDaWork'03 Beardah and others (2003) presented some analyses that suggested that, for some typical archaeometric glass compositional data, what has been called 'crude' principal component analysis (PCA) of standardized data Aitchison, 1996, p. 186) often appeared to produce more interpretable results than PCA of log-ratio transformed data (LRA). This is despite the fact that authors such as Aitchison and others (2002), who advocate LRA, have described the former approach as 'meaningless' and 'inappropriate'.

Notwithstanding this description, and the theoretical reasons that support it, it is rather easy to produce realistic examples (using both simulated and real data) where crude PCA produces archaeologically interpretable results much more readily than LRA. The issues involved, with examples, are explored in Baxter and others (2005), Baxter and Freestone (2005) and this paper. The term 'crude PCA' is used throught this paper to distinguish it from PCA of unstandardized log-ratio transformed data.

The fundamental practical problem in applying PCA to unstandardized log-ratio transformed data is the well-known difficulty that analysis will be dominated by those variables with the highest variance. For glass compositional data the variables with the largest variance on a log-ratio scale are typically those with low absolute values on the untransformed scale, and these are often, though not invariably, non-structure-carrying. Usually some form of variable selection is necessary before LRA 'succeeds' in identifying interpretable structure, which is often the same as that suggested by crude PCA, but not always as clear.

Technically, PCA is accomplished via an eigen-decomposition of the covariance or correlation matrix of the (possibly transformed) data, or via a singular value decomposition of the data matrix. A well-documented argument for not using crude PCA with compositional data is that covariances and correlations do not have a meaningful interpretation. We do not dispute this, but note that PCA can also be viewed simply as an unsupervised pattern-seeking method, the success of which is determined by whether or not interpretable results are consistently obtained, as judged by domain-specific considerations.

If there are p correlated variables, PCA tranforms the input data into p new uncorrelated variables that are linear combinations of the originals. The first principal component has maximum variance; the second has second maximum variance subject to the lack of correlation; and so on. As measured by variance, therefore, the PCs are ordered in terms of their importance. The usual expectation is that two- or three-dimensional plots based on the first few PCs will reveal any structure in the data and, where interpretable stucture exists, this is often the case for glass compositional data (with the caveat that obvious outliers may sometimes need to be stripped from an analysis if they dominate the appearance of a plot).

It was suggested at CoDaWork'03 that in our LRAs we might have been missing structure revealed by the 'less important' PCs. It is well-known that, in principle, this can happen and is easy enough to check by inspecting all possible pairwise PC plots. For the record, in none of the analyses reported in the references cited previously did LRA reveal structure using the less important components.

These issues are investigated further in the third section of the paper, after establishing terminology and notation in the next section.

2 Notation and terminology

Let n observations be available for p variables, $X_1, X_2, \ldots, X_p$, with observation i for variable j denoted by x_{ij}. The term *completely compositional* will be used for variables for which $\sum_j x_{ij} = 100\%$, and *sub-compositional* for a subset of such a set of variables. If the data are sub-compositional two approaches are possible. One is to convert the data to what Barceló-Vidal (2003) called *fully compositional* data, by defining a *residual* variable as $(100 - \sum_j x_{ij})$. The other is to normalize the variable sum to 100% and treat the normalized data as completely compositional.

In the LRA using PCA for completely compositional data recommended in Aitchison (1986), centered log-ratios of the form

$$y_{ij} = \log[x_{ij}/g(\mathbf{x}_i)]$$

are used, where

$$g(\mathbf{x}_i) = (x_{i1}x_{i2}\ldots x_{ip})^{1/p}$$

is the geometric mean of the data for the ith composition.

PCA of standardized data is based on

$$y_{ij} = (x_{ij} - \bar{x}_j)/s_j$$

where $\bar{x}_j$ and s_j are the mean and standard deviation of variable j.

3 Example

Freestone and others (2000), in a study of primary glass compositions from archaeological sites in Israel, distinguished between two types of glass termed *Levantine I* and *Levantine II*. The glass studied came from geographically separated sites, with the *Levantine II* glass also later. Since the original paper was published further analyses (some unpublished) have been added to the original data and it is the expanded data set that is used here. Close inspection of the data suggests that some cases may be misclassified, so here we treat the analysis as an unsupervised pattern recognition problem for illustration, rather than taking the classification as given.

A simple model for the composition of primary glass is to assume that glass (G) is made from two ingredients, sand (S_1) and a source of soda (S_2), combined in a recipe that is determined by a mixing proportion, π. Thus

$$G = \pi S_1 + (1 - \pi) S_2.$$

Analysis will focus on the sub-composition SiO_2, Al_2O_3, CaO, FeO and MgO, components of the sand assumed not to be present in the soda source. Of these components silica (SiO_2) is invariably dominant, with values in the rand 65.90% to 77.02% for our data. The components of the source of soda are not used in our analyses, but the dominant component, Na_2O ranges between 10.28% and 18.87%.

3.1 Crude PCA

A PCA of standardized data, omitting seven obvious outliers, results in Figure 1, and shows two clear compositional groups. For comparison with later figures the labelling into two groups, "x" and "o", is that *suggested by this figure*. The group to the right is very similar to, but not identical with, the originally defined *Levantine I*; similarly, the group to the left is close to, but not identical with, the originally defined *Levantine II*. Henceforth we use the terms *Levantine I* and *Levantine II* to refer to the two chemical compositional groups defined by this figure.

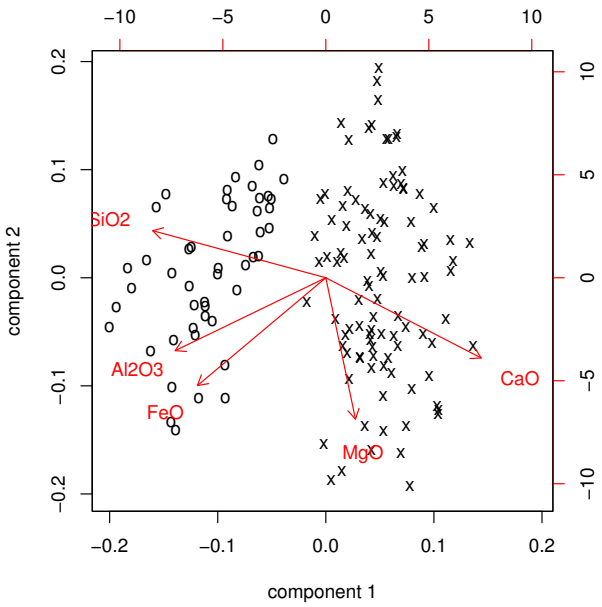

Figure 1: A biplot of the first two principal components of the Levantine sand compositional data, using standardized data. Seven compositional outliers have been omitted. See the text for a discussion of the labelling.

Notwithstanding the use of crude PCA, the results make good archaeological sense. In our simple

model the soda may be regarded as 'diluting' the sand composition (i.e. the more soda the less sand), so that even though it is not included in the statistical analysis it influences it because the dilution has not been corrected for. The later *Levantine II* glass has notably lower levels of soda and correspondingly higher levels of silica compared to *Levantine I* (for the groups in Figure 1 the means of silica and soda are 69.6% and 15.7% for *Levantine I* and 74.2% and 12.5% for *Levantine II*). The analysis of the standardized raw compositions is picking up these features of the data, which are associated with known geographical differences that may also be interpreted as reflecting a reduction in the availability of natron (the source of the soda) over time, and therefore an increase in silica.

3.2 LRA of fully compositional data

If a residual is defined so that the data are fully compositional, Figure 2 results. Separation of the two types previously defined is good, but by no means perfect, so that interpretation is equivocal.

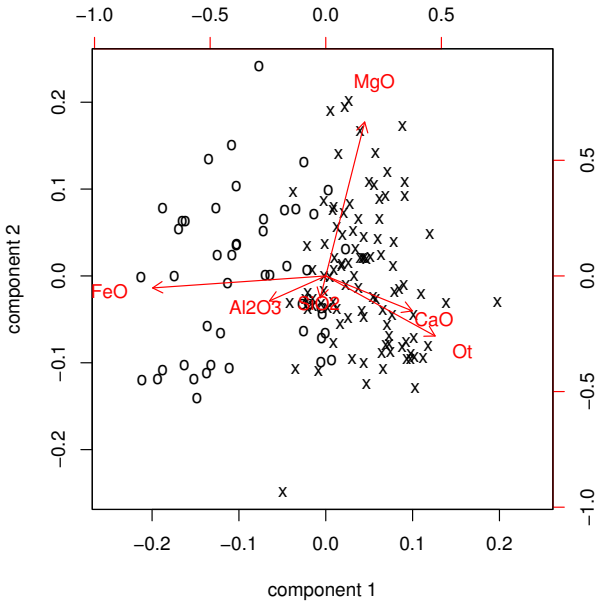

Figure 2: A biplot of the Levantine sand compositional data, using log-ratio analysis after converting to fully compositional data by defining a residual ("Ot" in the plot). *Levantine I* is labelled "x" and *Levantine II* "o". The seven compositional outliers omitted from the previous figure have also been omitted here.

The dominant variables in the biplot are MgO and FeO. Bearing in mind what has been said about such plots being determined by variables that may not be structure carrying, we investigate what happens if we merge these two variables with the residual. Figure 3, which now shows very good separation between the groups, results.

In Baxter and others (2005) it was suggested that where there is cluster structure in a set of compositional data, defined by differences in the absolute values of some variables, crude PCA would typically identify this structure. It was argued that with such structure it should be possible to select a subset of the variables which, with the residual defined as the difference between 100% and the sum of this subset, would lead to the definition of ratios that exhibited the same cluster structure. This is what appears to be happening here, so that the cluster structure revealed in both analyses is telling the same story. The crude PCA identifies the structure more directly.

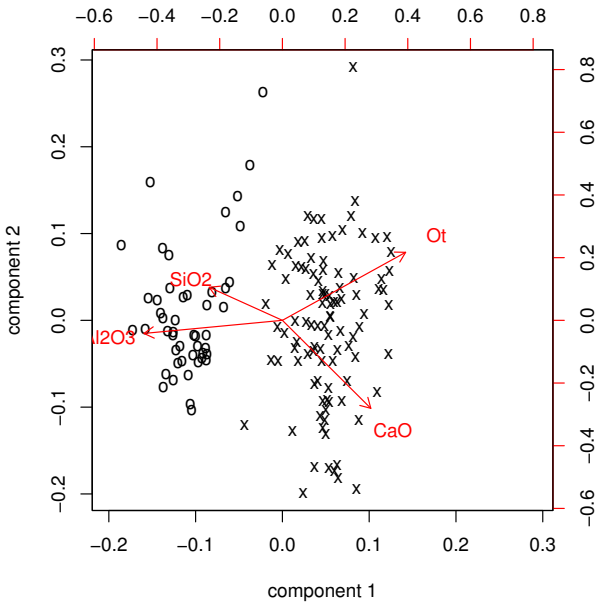

Figure 3: A biplot of the first two principal components of the Levantine sand compositional data, using log-ratio analysis after converting to fully compositional data by defining a residual and merging FeO and MgO with the residual. *Levantine I* is labelled"x" and *Levantine II* "o". The seven compositional outliers omitted from Figure 1 have also been omitted here.

3.3 LRA of normalized data

An apparently similar result, but in fact one telling a different story, emerges if LRA, normalizing the sum of the oxides to 100%, is used. This leads to the biplot of Figure 4. Here, *Levantine I* plots mostly to the left and *Levantine II* to the right, but there is also overlap. The dilution effect has been eliminated because of the use of relative values so that we are directly comparing the sands used for the two types. On a first interpretation Figure 4 might suggest that the two types cannot be associated unequivocally with distinct sources of sand. This is an archaeologically plausible interpretation and, taken in conjunction with the analysis of standardized data, would imply that while the recipe for glass making changed over time, the source of the raw materials did not.

As before, bearing in mind the fact that in LRA non-structure-carrying variables can obscure interpretable pattern in the data, what happens if the dominant variables, MgO and FeO, are removed from the analysis and the remaining variables renormalized? Figure 5 now results. This separates *Levantine I* and *Levantine II*, as in Figures 1 and 3, but has a different interpretation. Whereas the previous results could be interpreted as showing clusters associated with differences in the absolute values of the variables, and hence a difference in the recipes used (i.e. lower silica, higher soda versus higher silica, lower soda), here the clustering can be interpreted as suggesting a change in the source of sand for the two types (i.e different make-ups of the sand sources). This is more in accord with what we suspect from the archaeological evidence.

4 Discussion

To summarise, the crude PCA analysis and the fully compositional LRA analyses after variable selection, suggest the same groups in the data. These arise because of differences in the absolute values of some of the variables that can be interpreted in terms of a change in the recipe used to make the glass. It is typical, in our experience, that crude PCA recovers such structure rather

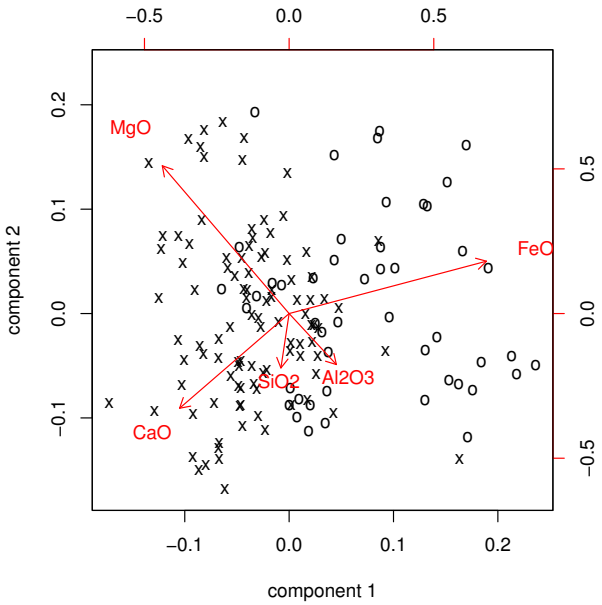

Figure 4: A biplot of the normalized Levantine sand compositional data, using log-ratio analysis. *Levantine I* is labelled "x" and *Levantine II* "o". The seven compositional outliers omitted from previous figures have also been omitted here.

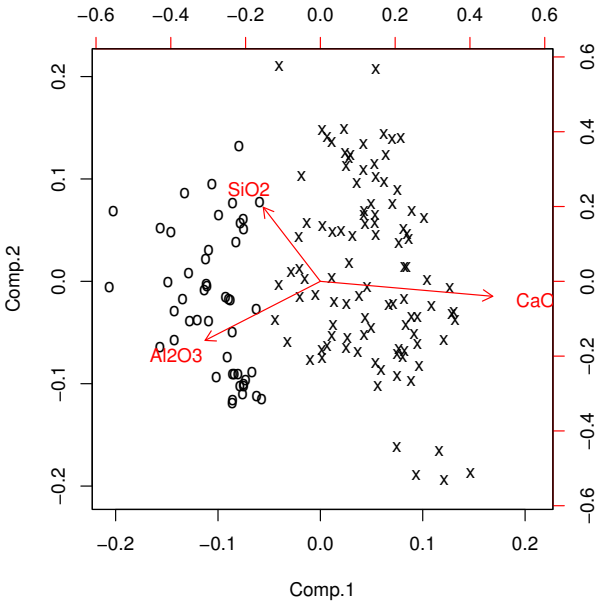

Figure 5: A biplot of the first two principal components of the normalized Levantine sand compositional data, using log-ratio analysis and omitting FeO and MgO. *Levantine I* is labelled"x" and *Levantine II* "o". The seven compositional outliers omitted from previous figures have also been omitted here.

directly. It is to be expected that LRA of fully compositional data, after suitable variable selection and definition of a residual, will also recover this structure, and this also has been our experience. It does so less directly than crude PCA. The correspondence between the two forms of analysis undermines claims that crude PCA produces 'meaningless' results, if what one is interested in is pattern recognition rather than the interpretation of covariance structure.

The LRA of normalized data produced superficially similar results, but in fact has a different interpretation. Archaeologically there is no doubt that the two types reflect geographically and chronologically distinct sites of production. Different recipes were used for glass-making, possibly because the supply of one of the ingredients that introduced the soda in the recipe began to 'dry-up'.

As far as the other ingredient, the sand, is concerned it is possible either that the same source continued to be used, or that the source as well as the recipe changed. The normalized LRA analysis bears directly on this question. The initial analysis produces results consistent with the first possibility but, after omitting MgO and FeO and renormalizing the data, results consistent with the second possibility are obtained. This, for archaeological reasons, is currently the preferred interpretation, though research continues.

An axiom of LRA seems to be that compositions provide information on relative values and that absolute diffences should not be of interest. The examples just described suggest that absolute differences can be of interest, and that pattern seeking methods such as PCA can recover interpretable structure based on the standardized compositional data. It seems to us indisputable that analysis of standardized data sometimes produces more archaeologically interpretable results than LRA, or does so with greater ease. The challenge for proponents of LRA is not to reassert the theoretical cachet of the methodology, but rather to demonstrate its practical utility when applied to typical archaeometric data and questions. In this context the issue of variable selection, and the fact that for a focused choice of variables the results of normalized LRA can complement those from crude PCA (or fully compositional LRA for those wanting to remain within an LRA framework), are problems meriting further investigation.

Acknowledgements

We are grateful to those participants at CoDaWork'03, whose comments on earlier work helped us articulate what we think some of the issues are; in particular, the paper presented by Carles Barceló-Vidal greatly helped to clarify some of our thinking.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.

Aitchison, J., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2002). Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis. *Archaeometry 44*(2), pp. 295–304.

Barceló-Vidal, C. (2003). When a data set can be considered compositional? *CoDaWork'03: Compositional Data Analysis Workshop, Girona, Spain*. (http://ima.udg.es/Activitats/CoDaWork03/)

Baxter, M.J. and Freestone, I.C. (2005). Log-ratio compositional data analysis in archaeometry. (Submitted for publication).

Baxter, M.J., Beardah, C.C., Cool, H.E.M. and Jackson, C.M. (2005). Compositional data analysis of some alkaline glasses. *Mathematical Geology*, *37*(2), pp. 183–196.

Beardah, C.C., Baxter, M.J., Cool, H.E.M. and Jackson, C.M. (2003). Compositional data analysis

of archaeological glass: problems and possible solutions, *CoDaWork'03: Compositional Data Analysis Workshop, Girona, Spain*.
(http://ima.udg.es/Activitats/CoDaWork03/paper_baxter_Beardah2.pdf)

Freestone, I.C., Gorin-Rosen, Y. and Hughes, M.J. (2000). Composition of primary glass from Israel. In M.-D. Nenna (Ed.), *La route du verre: ateliers primaires et secondaires de verriers du second millinaire av. J.-C. au Moyen-Age*, Travaux de la Maison de l'Orient Méditerranéen no. 33, pp. 65-84.